An Empirical Study of Translation Hypothesis Ensembling with Large Language Models

António Farinhas[1,2], José G. C. de Souza[3], André F. T. Martins[1,2,3]
[1]Instituto Superior Técnico (Lisbon ELLIS Unit), [2]Instituto de Telecomunicações, [3]Unbabel

There's lots of research on task-specific NMT models but LLMs offer a new perspective!

We generate multiple hypotheses by using a single prompt and sampling multiple times; we ensemble them using different techniques.

Translation Hypothesis Ensembling

Using external quality estimation/evaluation models

- ranking with QE: $\hat{y}_{\text{ranking}} = \text{argmax}_{y \in \bar{\mathcal{Y}}} \text{CometKiwi}(y)$
- MBR decoding: $\hat{y}_{\text{mbr}} = \text{argmax}_{y \in \bar{\mathcal{Y}}} \mathbb{E}_{Y \sim p_\theta}[\text{COMET}(Y, y)]$

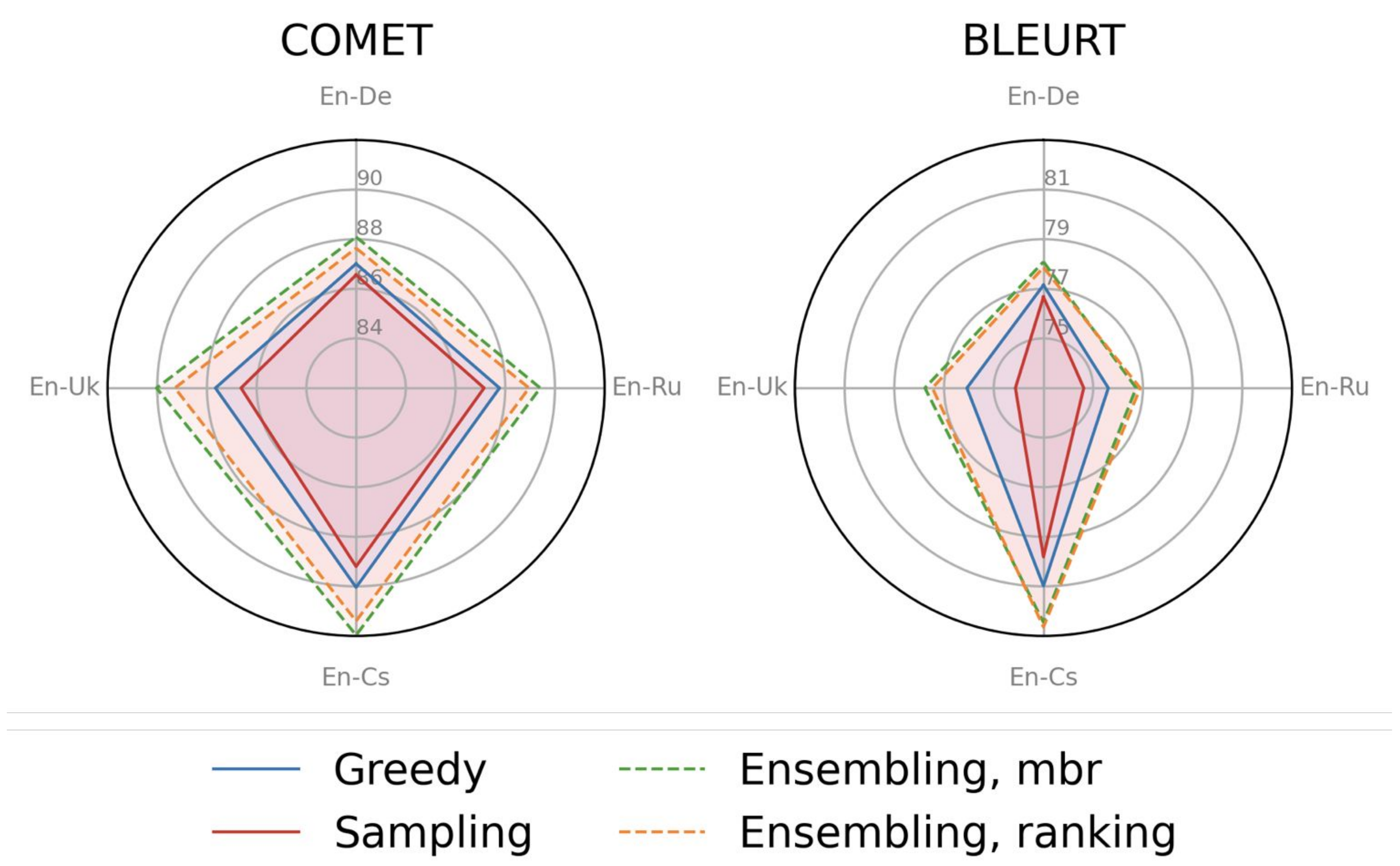
Using the LLM

- ChooseBest: formulated as a multiple choice question
- GenerateBest: asking the LLM to generate a final prediction

Check the paper for more LPs and analysis!

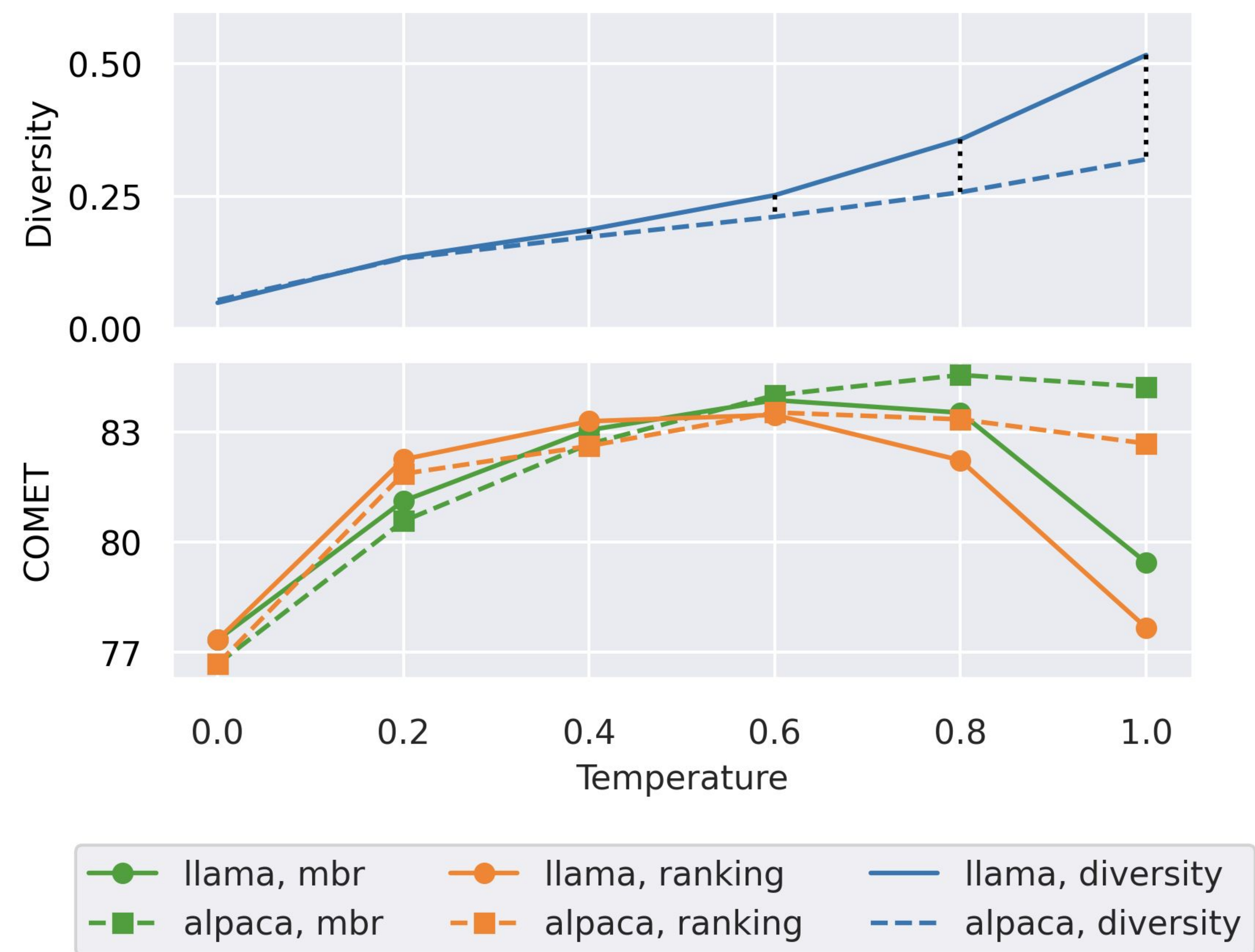
ChatGPT

- translation quality can be enhanced with a small number of unbiased samples, especially for EN-X



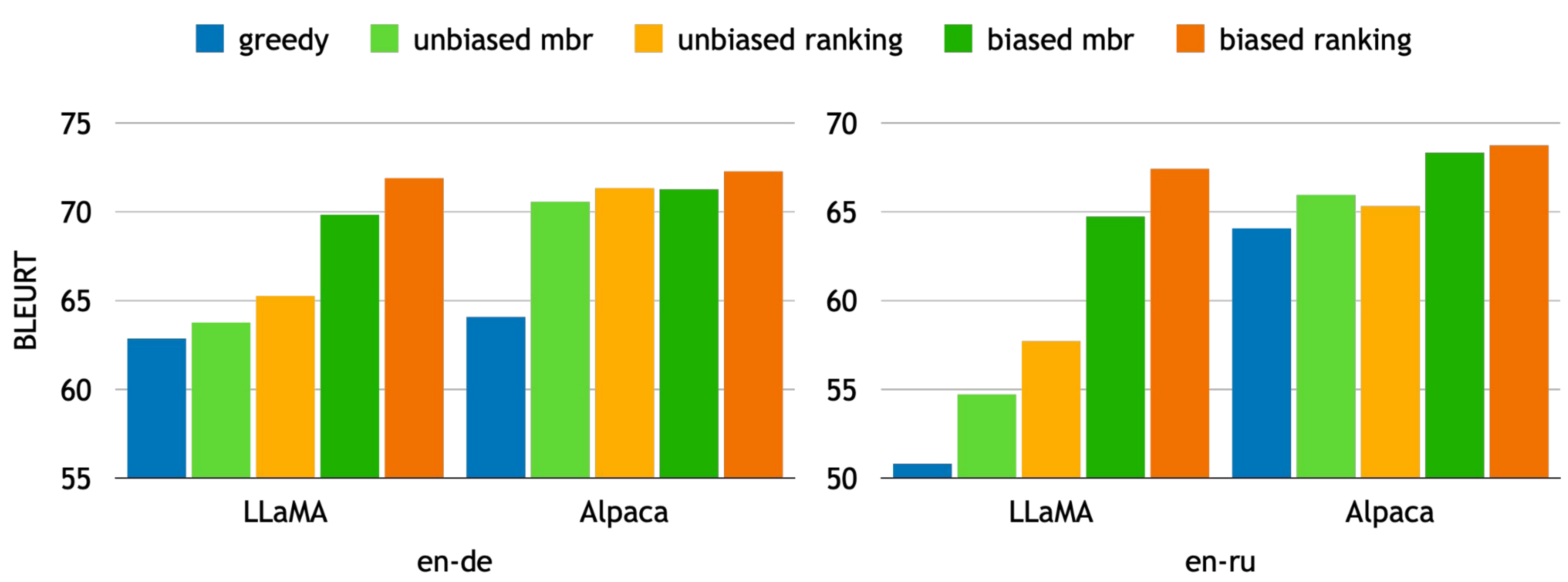
Biasedness, diversity, and quality

- the diversity between hypotheses increases with the sampling temperature at a different rate for LLaMA and Alpaca



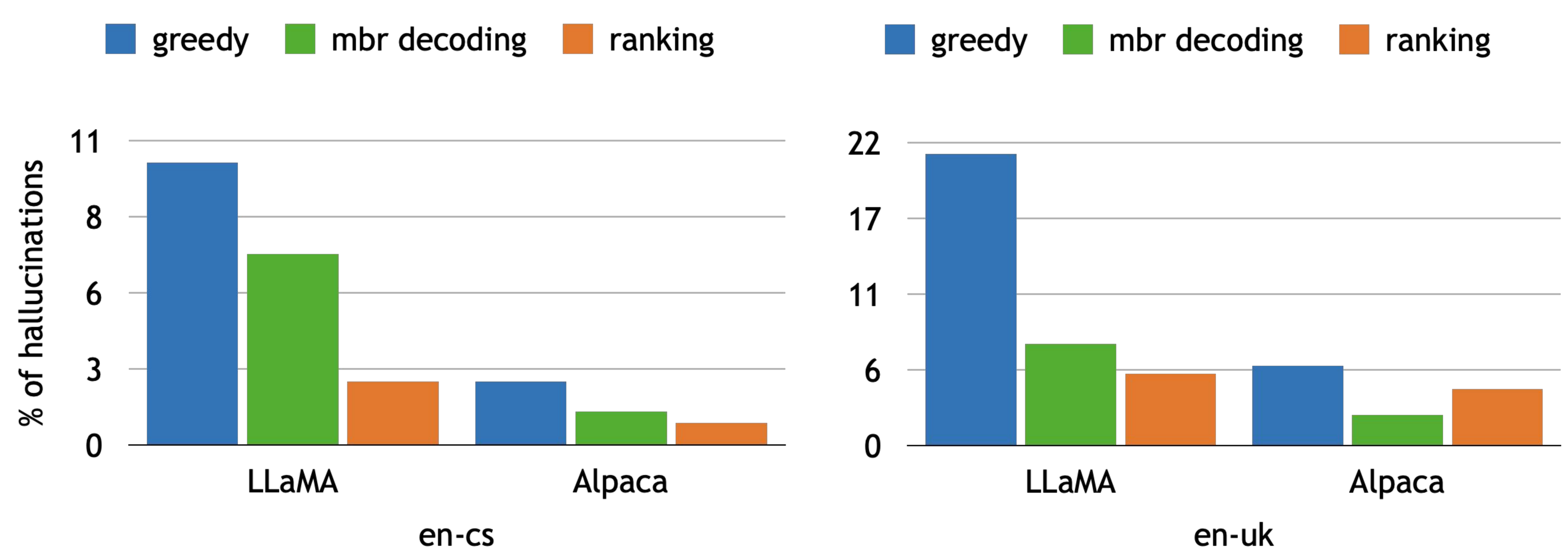
LLaMA and Alpaca

- ensembles of unbiased samples from LLaMA don't perform well
- alpaca performs better and biasing samples boosts performance



Hallucinations

- hallucination rate decreases with instruction tuning
- ensembling translations decreases the number of hallucinations



Acknowledgments: EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), DECOLLAGE (ERC-2022-CoG 101088763), Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), and Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.